

# Συνοψεις για Δεδομένα XML με Ετερογενές Περιεχόμενο



Άλκης Πολυζώτης  
*UC Santa Cruz*  
Μίνως Γαροφαλάκης  
*Intel Research, Berkeley*

## Ανακεφαλαίωση

- Η συνοψηση είναι σημαντικό κομμάτι της διαχείρισης δεδομένων
  - ♦ Βελτιστοποίηση ερωτήσεων
  - ♦ Προσεγγιστικές απαντήσεις
  - ♦ Ανεύρεση προτύπων
- Συνοψεις VTreeSketch
  - ♦ Δομή + Τιμές
  - ♦ Ετερογενές περιεχόμενο

# Συνόπιση XML Δεδομένων

QuickTime™ and a  
TIFF (uncompressed) decompressor  
are needed to see this picture.

Αρκετά KBs

XML  
Synopsis

Ερώτηση Q

Αποτέλεσμα R'

Προσεγγίζει

XML  
Data

Ερώτηση Q

Αποτέλεσμα R

Αρκετά MBs

R' υπολογίζεται πιο γρήγορα!

# Εφαρμογή: Εκτίμηση Επιλεκτικότητας

QuickTime™ and a  
TIFF (uncompressed) decompressor  
are needed to see this picture.

- Η βελτιστοποίηση βασίζεται σε **παράγοντες επιλεκτικότητας**
  - ♦ Π.χ.: //author[name="Tova"]/paper θα χρειαστεί //author, //name, //paper, //author/paper, ...
- Οι ακριβείς τιμές είναι μη πρακτικές => Εκτίμηση!

XML  
Synopsis

COUNT(Q)

Επιλεκτικότητα S'

XML  
Data

COUNT(Q)

Επιλεκτικότητα S

## Παράδειγμα

- Βιβλιογραφικά δεδομένα

```
<entry>  
  <year>1996</year>  
  <author>N.Alon</author>  
  <author>Y.Matias</author>  
  <author>M.Szegedy</author>  
  <title>The space complexity of approximating the frequency  
  moments  
</title>  
  <abstract>...</abstract>  
</entry>
```

- Ερώτηση

```
entry[year>2000][author="Matias"]/abstract[ftcontains stream data]
```

## Δυσκολίες/Προκλήσεις

QuickTime™ and a  
TIFF (Uncompressed) decompressor  
are needed to see this picture.

- Ετερογένεια περιεχομένου
  - ♦ Δενδρική δομή, αριθμητικές τιμές, αλφαριθμητικά, κείμενο
- Ετερογένεια συνθηκών
  - ♦ Δομικές, εύρους, sub-string, term queries
- Συσχέτιση δομής/τιμών
- Κατανομή χώρου συνόψισης μεταξύ:
  - ♦ Δομής/Τιμών
  - ♦ Τιμών διαφορετικών τύπων

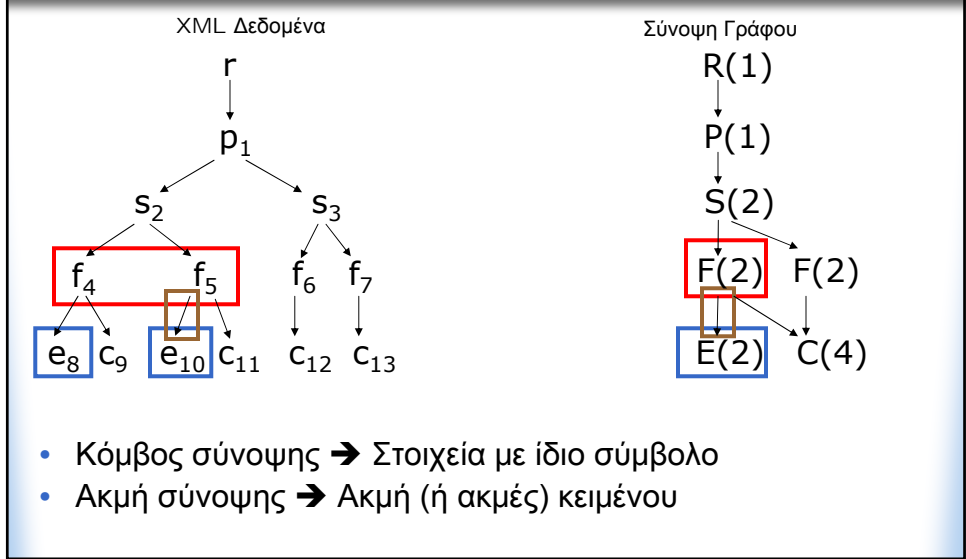
## Λύση: VTreeSketch (XCluster)

- Συνόψεις για δενδρικά XML δεδομένα
  - ♦ Δομή + Τιμές
  - ♦ Τιμές διαφορετικών τύπων
- Προσεγγιστικές απαντήσεις σε δενδρικές ερωτήσεις με ετερογενείς συνθήκες
- Αποδοτική κατασκευή
- Υψηλή ποιότητα συνοψισης με χαμηλές απαιτήσεις χώρου

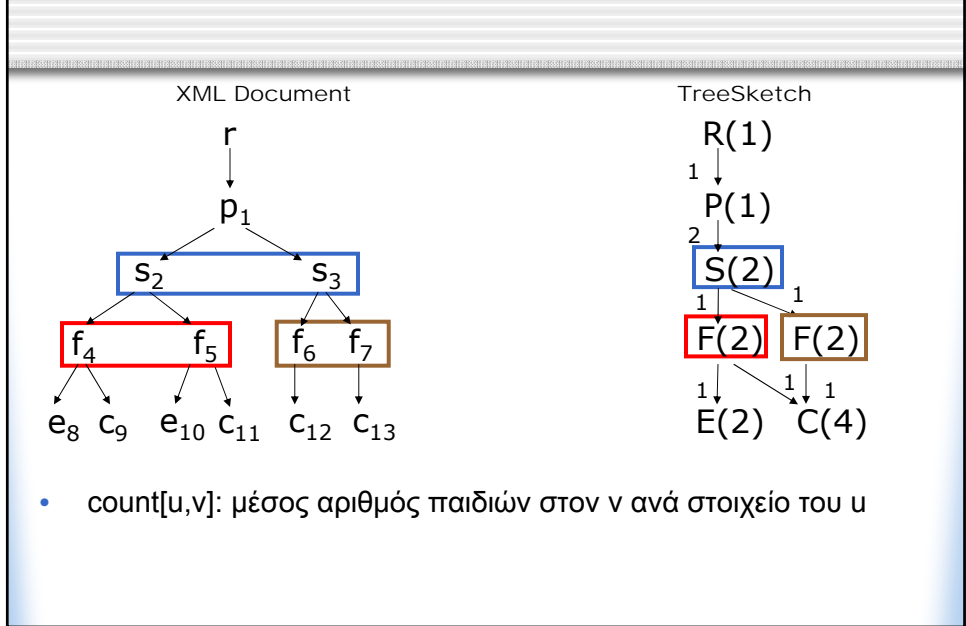
## Διάγραμμα Παρουσίασης

- Μοντέλο Συνοψισης
- Αλγόριθμος Κατασκευής
- Πειραματικά Αποτελέσματα
- Επόμενα Βήματα

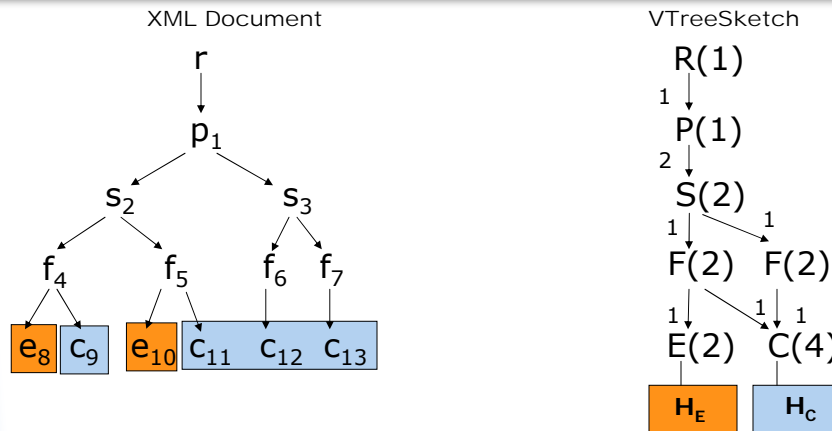
# VTreeSketch: Πληροφορία Γράφου



# VTreeSketch: Πληροφορία Δομής



## VTreeSketch: Πληροφορία Τιμών

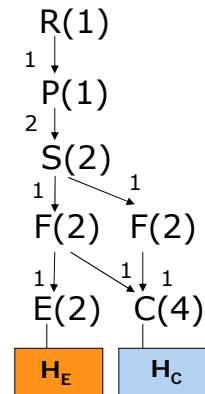


## Συνοψεις Τιμών

- Μονοδιάστατες συνοψεις
- Υλοποίηση εξαρτάται από τον τύπο των τιμών
  - ♦ Αριθμητικές τιμές: ιστογράμματα
  - ♦ Αλφαριθμητικά: παραλλαγή Pruned Suffix Tries
  - ♦ Κείμενο: end-biased term histograms
    - Συνδυασμός ιστογραμμάτων και bitmap indices

## Συνόψιση = Συσταδοποίηση

- Κόμβος  $\Leftrightarrow$  Συστάδα δομής/τιμών
- Συνοχή εξαρτάται από:
  - ♦ Ομοιότητα δομής
  - ♦ Ομοιότητα κατανομής τιμών
- Καλή συνοχή  $\rightarrow$  Ακριβής σύνοψη
- Κύρια δυσκολία: ετερογένεια!

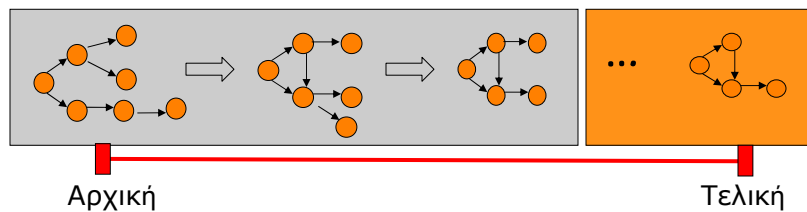


## Κατασκευή Συνοψεων

- Ζητούμενο: κατασκευή **αποδοτικής σύνοψης** για συγκεκριμένα **δεδομένα T** και για περιορισμένο **χώρο αποθήκευσης B**
- Πρόβλημα συσταδοποίησης... αλλά με αυξημένη δυσκολία!

## Αλγόριθμος Κατασκευής

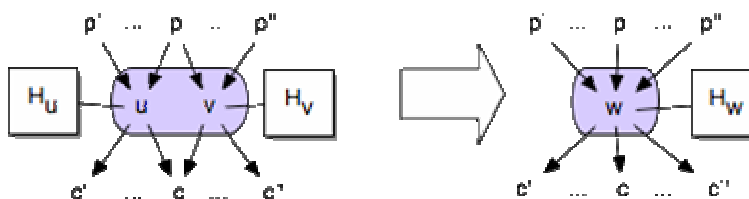
- Σταδιακή συμπίεση λεπτομερούς σύνοψης
  - ♦ Φάση A: Συμπίεση δομής
  - ♦ Φάση B: Συμπίεση κατανομών
- Επιλογή βημάτων με βάση την “απόσταση” μεταξύ αρχικής και τελικής σύνοψης



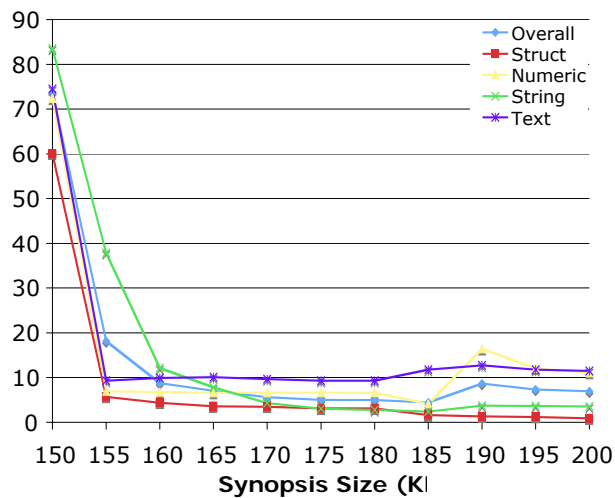
## Συνάρτηση Απόστασης

- Υπολογισμός βάσει “τοπικών” μ-ερωτήσεων

$$|u| \cdot \sum_{p,c} (\text{count}(u[p]/c) - \text{count}(w[p]/c))^2 + |v| \cdot \sum_{p,c} (\text{count}(v[p]/c) - \text{count}(w[p]/c))^2$$



## Πειραματικά Αποτελέσματα



- IMDB
  - ♦ Μέγεθος ~ 7MB
  - ♦ 236822 στοιχεία
- Μονοπάτια τιμών
  - ♦ 2 αριθμητικά
  - ♦ 4 string
  - ♦ 1 κείμενο
- 50KB πληροφορία δομής

## Επόμενα Βήματα



- Συνόψεις κειμένου και αλφαριθμητικών
- Αποδοτικότερος αλγόριθμος κατασκευής
- Εφαρμογή σε σχεσιακά δεδομένα

